Marek Gajęcki*

INFORMATION RETRIEVAL USING ASSOCIATION LISTS

The paper presents the main models of information retrieval in text. The idea of association list and its use for information retrieval is introduced. The paper contains the evaluation results of the algorithms under discussion.

Keywords: *Natural Language Processing, Statistical Inference, Information Retrieval*

ZASTOSOWANIE LIST SKOJARZENIOWYCH W WYSZUKIWANIU INFORMACJI

W artykule zaprezentowano podstawowe modele wyszukiwania informacji w tekście. Przedstawiono ideę listy skojarzeniowej i jej zastosowanie w wyszukiwaniu informacji. Zawarto rezultaty zawierające ocenę jakości prezentowanych algorytmów.

Słowa kluczowe: *przetwarzanie języka naturalnego, wnioskowanie statystyczne, wyszukiwanie informacji*

1. Introduction

The mankind entered the 21st century as a highly advanced information society. Fast progress in science combined with the emergence and expansion of information and telecommunication technologies has caused that information has become a main production resource of human. On the one hand, information is available for everybody, on the other hand, an enormous increase in the amount of electronic documents makes accessing this information very difficult or even impossible. One of the possible ways out of this situation is to provide effective tools for information retrieval, which enable to find relevant information.

The rest of the paper presents the main models of information retrieval in a repository of textual documents, describes the concept of *association list* and its use for improving the quality of retrieval. Moreover, it compares the models with an example of information retrieval in a repository of press notes.

*Institute of Computer Sciences, AGH University of Science and Technology, Kraków, Poland, mag@agh.edu.pl

2. Association List

Association list is a structure that groups the words which are thematically connected with a given word. The construct of a list is based on an assumption that a human's statement maps the system of concepts with which the speaker operates [1]. If we can access a relatively large number of statements (as texts), based on the statistical inference, we can create a list of words associated with a given word. Since the concepts in a text are connected with nouns, the defined words as well as the defining word on a association list will be represented by nouns.

Association list does not distinguish between the kinds of connections between words, but their existence and strength of connections. The strength of connections is evaluated based on statistical parameters like the number of occurrences of a word in a corpus of texts or the number of occurrences of a defined word and defining word. The algorithms for generating association lists are described in [2]. An association list is the first stage to build a semantic vocabulary similar to the WordNet dictionary [3].

An example of association list for the word *student* is shown below (We place the top ten elements of the list). The related words used are *uczelnia* ('higher school'), *uniwersytet* ('university'), *szkoła* ('school'), *uczeń* ('student'), *politechnika* ('technical university'), *studium* ('college'), *studio* ('studio'), *zrzeszenie* ('association'), *wydział* ('faculty'), and *stypendium* ('scholarship').

uczelnia	95	648	14.6	179.3
uniwersytet	79	1350	5.8	123.9
szkoła	63	2343	2.6	83.7
uczeń	40	618	6.4	76.3
politechnika	26	173	15.0	63.9
studium	27	382	7.1	58.2
studio	27	408	6.7	57.4
zrzeszenie	17	53	33.0	52.0
wydział	24	732	3.2	44.0
stypendium	17	145	11.7	43.1

The consecutive columns contain: a defining noun, the number of occurrences of the defining noun along with the defined noun, the total number of occurrences of the defining noun in the corpus, the percentage ratio of the latter two ones, the strength of connections.

3. Models of Information Retrieval

3.1. Boolean Queries Model

The Boolean Queries model is a simpler method of information retrieval in a set of textual documents. A query to the system which implements this model is *(boolean query)*. A boolean query is built of terms connected by conjunctions like AND, OR,

NOT. Terms correspond to words and phrases. A term is considered true, if the word occurs in the text. The procedure of information retrieval in the Boolean Queries Model consists in selecting such items from document repositories, for which the query is true.

In case of an inflection language, e.g. Polish, the Boolean Queries Model should be expanded. It must take into account that in Polish a single word is represented by many inflection forms. It implies the necessity to use a tool like an inflection dictionary [4, 5]. This dictionary enables to expand a query so to take into account the inflection forms of a word. E.g. a query:

`student AND egzamin` ('student AND exam')

will be extended to a form which takes into account the inflection features of the language

```
(student OR studenta OR studentowi OR studentem OR studencie OR studenci
OR studentów OR studentom OR studentami OR studentach) AND (egzamin OR
egzaminu OR egzaminowi OR egzaminem OR egzaminie OR egzaminy OR egzaminów
OR egzaminom OR egzaminami OR egzaminach)
```

In the above example, we can see that the use of the inflection dictionary enables to find documents which contain any inflection form of the word.

3.2. Vector Space Model

In the Vector Space Model, texts are represented by the vectors of factors assigned to the words occurring within the text. In this model a query results in a vector form by considering it as a textual document. From many methods intended to determine the factors of the vector which corresponds to a particular document and query, the most frequently used method is to choose a value dependent on the number of occurrences of a word in the document and on the number of occurrences of the word in document repositories.

$$weight = TF(w,d) \cdot \log(N/DF(w)) \tag{1}$$

where:

TF – number of occurrences of a word w in the document d (*term factor*),

DF – number of the documents where a word w occurs (*document factor*),

N – number of documents.

The main task of search in the Vector Space Model is to determine a distance (or similarity) between a query and a document. This distance is a measure of semantic proximity between two texts.

The most popular method to determine the distance (similarity) is (*cosine measure*) that consists in computing the cosine of the angle between the vectors which represent the texts.

$$\cos(\alpha) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2} \cdot \sqrt{\sum y_i^2}} \tag{2}$$

An important advantage of this measure is its independence of the document size. There are also other measures: Euclides measure, Dic's measure, Jaccard's measure [6].

Independently of the selected method for measuring the distance or similarity between vectors, some value is obtained. This enables to rank documents according to their relevance to a query. This results in a ranking which allows to reduce the number of the returned results of a search.

3.3. Vector Space Model with Association List

A main drawback of the basic Vector Space Model is the inconvenient method for building a query. In this case a query can be a textual document, for which similar documents are searched for, or be a conjunction query. A more convenient method is the second one, which needs to provide a few words only. On the other hand, a short query may be disadvantageous for computing the distance between the query and documents.

To avoid problems of this kind, a *query expansion* is used. A query can be expanded, e.g. via using a relevant dictionary which contains semantic connections between words, for example synonyms. Since for the Polish language there is no dictionary similar to WordNet [3], so in order to expand the query, the above association lists are used.

Expanding a query with association lists is realized as follows: for each noun from within a query, a list of nouns associated with this noun is generated, then the top ten elements of each list are entered into the query. The procedure of constructing a vector for a query consists in assigning to words their weights resulting from the strength of connections between the words within association lists. In the resulting vector, the word weights are proportional to the strength of connections on the association list.

An example of a vector construct resulting from a query:

`student AND egzamin` ('student AND exam')

is given below:

```
[(student,0.58),(exam,0.58),(maturity,0.41),(test,0.23),
(mathematics,0.17), (High School Diploma,0.13),(technical university,0.09)
(higher school,0.08),...]}
```

In addition to a better form of vectors for queries, this solution takes into account the fact that quite often the user cannot define his/her information needs by a query.

4. Evaluation measures of effectiveness

To compare the effectiveness of retrieval methods we used three main measures: *precision, recall, and fallout* [7].

Precision determines the extent to which the retrieved documents match the query.

$$precision = \frac{number\ of\ returned\ relevant\ items}{number\ of\ returned\ items} \tag{3}$$

$$precision = \frac{number\ of\ returned\ relevant\ items}{number\ of\ returned\ items} \tag{4}$$

Recall is connected with the retrieval precision and determines the extent of success in retrieving all the relevant documents.

$$recall = \frac{number\ of\ returned\ relevant\ items}{number\ of\ all\ relevant\ documents} \tag{5}$$

Fallout determines the extent of return of nonrelevant items.

$$fallout = \frac{number\ of\ returned\ nonrelevant\ items}{number\ of\ all\ nonrelevant\ documents} \tag{6}$$

To enable determining the above parameters, it is necessary to manually assign to each item some information on its relevance to a given query.

5. Results

A comparison of retrieval methods was carried out on a repository of press notes of the Polish Press Agency (PAP) [9].

The first test consists in information retrieval from a set of 100 items, where 13 items relate to the stock exchange. The results of a retrieval for query `giełda` ('stock exchange') are shown in Table 1.

Table 1

Evaluation measures of information retrieval on "stock exchange"

Model	Boolean	Boolean + dictionary	Vector Space	Vector Space + list
Number of returned items	1	9	9 (9)	13 (26)
Number of relevant items	1	7	7 (7)	11 (12)
precision	100%	77.8%	77.8%	84.6%
recall	7.7%	53.8%	53.8%	84.6%
fallout	0.0%	2.3%	2.3%	2.3%

Since the vector space models return almost all documents (except documents which similarity is equal 0) we asume that number of returned items is equal to number of all relevant documents in the repository. This enable to express precision and recall as range from 0 to 100%.

The second test consisted in submitting a query `''student AND exam''` on a repository of 41182 items. Multiple trials resulted in 72 relevant items. The results are given in Table 2.

Table 2

Evaluation measures of information retrieval in a large document repository

Model	Boolean	Boolean + dictionary	Vector Space	Vector Space + list
Number of returned items	4	117	72 (2243)	72 (13589)
Number of relevant items	1	28	30 (71)	38 (72)
precision	25%	23.9%	41.6%	52.8%
recall	1.4%	38.9%	41.6%	52.8%
fallout	0.01%	0.2%	0.1%	0.08%

6. Conclusions

The research carried out so far allows to draw the following conclusions.

- To achieve satisfactory results in information retrieval in texts in Polish or other inflection languages, it is necessary to use an inflection dictionary of the given language. This substantially improves the recall of retrieved information at the cost of a slight decrease in precision.
- The Vector Space Model of information retrieval and the Boolean Query Model along with an inflection dictionary give qualitatively similar results.
- The use of association lists improves the precision and recall of information retrieval.
- Due to a significant time needed to construct an association list, it is reasonable to carry out the procedure of constructing lists and generation of an association dictionary [8].

7. Acknowledgements

The research reported here was partially supported by the KBN grant no. 3T11C05926.

References

[1] Palermo D. S., Jenkins J. J.: *Word Association Nouns: Grade School trough College*. 1st ed., Univeristy of Minnesota Press, 1964

[2] Lubaszewski W., Gajęcki M.: *Automatyczna ekstrakcja powiązań semantycznych z tekstu polskiego*. Computer Science, vol. 4. 2002, s. 119–130.

[3] Fellbaum Ch. (ed.): *WordNet. An Electronic Lexical Database*. MIT Press 1998, ISBN 0-262-06197-X.

[4] Lubaszewski W., Wróbel H., Gajęcki M., Moskal B., Orzechowska A., Pietras P., Pisarek P., Rokicka T.: *Słownik fleksyjny języka polskiego*. wyd. 1., Kraków, Wydawnictwo Prawnicze LexisNexis, 2001.

[5] Gajęcki M.: *Serwer leksykalny języka polskiego*. Computer Science, vol. 3, 2001, s. 131–150.

[6] Manning C. D., Schütze H.: *Foundations of Statistical Natural Language Processing*, 1st ed., MIT Press Cambridge, 2000, ISBN 0-262-13360-1.

[7] Rijsbergen C. J.: *Information Retrieval*. 2nd ed., London, Butterworths, 1979, ISBN 0408709294.

[8] Gajęcki M.: *Automatyczne generowanie słownika asocjacyjnego na podstawie korpusu tekstów*, V Krajowa Konferencja Naukowa, Inżynieria Wiedzy i Systemy Ekspertowe, Wrocław, 2003.

[9] `http://dziennik.pap.com.pl`.